MC2024
PEOPLE & BUILDINGS
Proceedings of 13th Masters Conference:
London, 16th September 2024

Network for Comfort &
Energy Use in Buildings
http://www.nceub.org.uk

Evaluating Large Language Models (LLMs)' Understanding of 'Black Crusts' Predictive Models for Built Heritage Preservation

Wenrui Sun[1], Dr. Josep Grau-Bove[2], Dr. Daniela Reggio[2]
[1] MSc Built Environment: Sustainable Heritage (Data Science), University College London, UK, Email: wenrui.sun.23@ucl.ac.uk;
[2] Institute for Sustainable Heritage, Bartlett School of Environment, Energy and Resources

Abstract: This study evaluates the understanding of LLMs in predicting limestone sulphation, described in common language as 'black crusts', which is an environmental decay damaging for historic buildings, implying gypsum formation on the surface of carbonatic materials.

The research question is: 'To what extent can built heritage managers use LLMs for preservation advice?'. GPT-4, Claude, and OpenArt were used for evaluation, and prompts were designed to test different aspects of limestone sulphation predictions. Gypsum thickness calculated at various time intervals from published studies was prompted to the LLMs to generate new predictions. These results were then cross-referenced with predictive models to assess accuracy.

The findings indicate that LLMs produced varying results each time, with significant discrepancies compared to published models. Numerical predictions, data fitting, and image forecasting based on LLMs were explored, underscoring the limitations of LLMs in predictive modelling. Further testing is required to leverage LLMs capabilities in heritage preservation.

Keywords: Limestone Sulphation, LLMs, Black Crusts, Heritage Preservation

1. Introduction

Limestone sulphation is a critical chemical weathering reaction where sulphur dioxide (SO_2) from combusted fuel pollutants in the atmosphere reacts with calcium carbonate ($CaCO_3$) in limestone from building materials forming calcium sulphate that in the atmosphere forms gypsum ($CaSO_4 \cdot 2H_2O$)(Giavarini et al., 2008). The latter is highly hygroscopic, and this reaction is particularly damaging in the context of preserving cultural heritage, as many historic buildings, monuments, and sculptures are composed of limestone and are vulnerable to degradation due to sulphation and formation of 'black crust'(Sabbioni et al., 1998). This degradation accelerates structural wear and alters the visual and physical integrity of historic stonework. Therefore, understanding limestone sulphation is important for developing effective preservation strategies.

The objective is to evaluate the effectiveness of Large Language Models (LLMs) in predicting limestone sulphation processes compared to the traditional mathematic models (Alì et al., 2007a, 2007b; Giavarini et al., 2008; Tran Thi Ngoc Lan et al., 2005). GPT-4, Claude and OpenArt were used for evaluation, and a variety of textual, visual and numerical prompts were designed to test different aspects of limestone sulphation predictions. Variables as gypsum thickness were calculated and measured at various time intervals from published studies was prompted to the LLMs systematically to generate predictions. These results were then cross-referenced with predictive models to assess their accuracy.

The research question addressed is: To what extent can built heritage managers use LLMs as tools for preservation advice, particularly in predicting the sulphation process of limestone? Nonetheless, the findings indicate that LLMs produced varying results each time, with significant discrepancies compared to experimental and numerical models, highlighting the limitations of LLMs in predictive modelling.

2. Methodology

2.1 Data Collection

Giavarini's research developed a mathematical model describing the diffusion and chemical reaction of SO_2 in calcium carbonate stones, predicting the progression of sulphation in porous stones like limestone. The model's validity was demonstrated through experimental methods and numerical simulations.

Experimental data for validation in the mentioned paper were collected using Carrara marble with 0.6% porosity, exposed to a pure SO_2 gas stream at 1.8 bar pressure, 100% relative humidity, and room temperature. Gypsum thickness was measured using Scanning Electron Microscopy (SEM), and concentrations of gypsum were analysed using Thermogravimetric Analysis (TGA). These data were compared with the results generated by the LLMs.

2.2 LLMs Selection

The following LLMs were selected for evaluation: GPT-4 by OpenAI, Claude by Anthropic and OpenArt. GPT-4 and Claude are specialised in generating nuanced text outputs, tailored for complex language tasks, while OpenArt focuses on creating image results.

2.3 Prompt Design

Prompting means supplying an LLM with preliminary information or context before asking the primary question you want answered.(Giray, 2023) A set of prompts are designed to test different aspects of the limestone sulfation process. First is to ask about the specific value predictions of different conditions. When giving commands, ask different LLMs: "How thick is the layer of gypsum formed on Carrara marble after X hours of exposure at 100% relative humidity, room temperature, pressure of 1.8 bar, and concentration of sulphur dioxide 0.1 ml/mol?" Secondly, instruct the LLMs to provide a model that explains the reasoning process of the sulfurization, specifically: "Provide the mathematical function that describes the rate of gypsum formation on the surface of limestone over time during the sulphation process.

2.4 Response Collection

To assess ChatGPT-4 and Claude's predictive capabilities, prompts were modified to include specific times: 24, 48, 72, 96, 120, 144, and 168 hours. The same prompt predicting gypsum thickness at 24 hours was run ten times to evaluate consistency. Data on the LLMs' performance in simulating gypsum thickness over time was collected. After generating initial data, experimental data was input to create fitted models. These prompts were re-entered to see if real experimental data improved the LLMs' accuracy and reliability.

2.5 Comparison and Analysis

The analysis focuses on accuracy under various conditions. An internal comparison is conducted, followed by a comparison between ChatGPT-4 and Claude to identify differences. Finally, LLM outputs are compared with established models in the literature.

2.6 Visual Predictions of Black Crust Formation Using OpenArt

In addition, image predictions using OpenArt were also implemented. To explore the visual impact of limestone sulphation on different buildings, architectural images were selected and uploaded to the OpenArt interface. Each image was processed with the prompt, "Draw the black crust formation due to limestone sulphation on this building." This task was designed to generate images that visually represent the effects of black crust development on diverse architectural surfaces over time. To assess the consistency and reliability, the same prompt was repeatedly inputted for each building image.

3. Results

3.1 Model Predictions and Experimental Data

Extracting the predicted data and experimental data from published models yields the following results. In the model predictions, gypsum thickness increases linearly with the square root of reaction time. The function describing this relationship is: Crust Thickness(um)= $56.357\sqrt{t}$ (h). This formula allows for the calculation by excel of gypsum thickness at any specified reaction time.

In the experimental validation conducted in a real-world setting, the thickness of the gypsum layer on Carrara marble at various reaction times was measured using SEM. The data are presented in the Table 1 below(Giavarini et al., 2008).

Table 1 Gypsum thickness measured by SEM at different reaction time

Time(h)	0	24	48	72	96	120	144	168
Thickness (um)	0	214	357	450	506	620	700	788

3.2 ChatGPT-4's predictions on limestone sulfation

Using the prompt, GPT-4 details reaction steps like chemistry equation, SO_2 concentration, and gypsum formation. However, it limits its analysis to basic chemical reactions, excluding crucial physical and biological factors for understanding limestone sulphation. GPT-4 predicts increasing crust thickness but significantly underestimates compared to validated data, as is shown in Figure 1. Repeated prompts for a 24-hour exposure showed high variability, with each run producing different numerical results, even with inconsistent units converted to micrometres (µm). Table 2 shows significant discrepancies from experimental data, indicating low accuracy and highlighting the model's limitations.

Table 2 Gypsum thickness predicted by ChatGPT-4 at 24h in 10 times run

Trial Times	1	2	3	4	5	6	7	8	9	10
Thickness (um)	7.42	74	5400	0.0054	0.1	5.4	0.001	0.55	0.53	0.0005
Accuracy	2.7%	26.8%	-1756.5%	0.002%	0.036%	1.96%	0.0004%	0.20%	0.19%	0.00018%

3.3 Model generated by ChatGPT-4

In ChatGPT-4, the prompt "Generate a predictive model for limestone sulphation..." was used. It consistently required experimental data to fit model parameters. By inputting experimental data from published models, a fitted predictive model was obtained. Although the fitting curve appears accurate, the quality of images and captions could improve. ChatGPT struggles to capture the linear relationship between gypsum thickness and the square root of time accurately. After 1000 hours, the GPT-4 model predicts a plateau in gypsum thickness, contrasting with the continual rise predicted by the validated model, as is shown in Figure 2.

3.4 Claude's predictions on limestone sulfation

Claude was found to be very cautious in its predictions. Unlike GPT-4, which provides specific figures, Claude responded, "Without detailed experimental results or a well-established model for this specific scenario, it's not possible to provide an exact thickness." It then outlined the factors that influence the process, including "Reaction kinetics," "Temperature," "Pressure," "SO2 Concentration," "Relative Humidity," and "Air flow." Upon further questioning for a predictive model, Claude was able to offer a simple model and provide numerical values. However, these results still showed significant differences when compared to published data. Without inputting a known model, Claude predicted that the thickness after 24 hours would be 0.823, which represents a hundredfold error compared to established results(Giavarini et al., 2008)

3.5 OpenArt's predictions on limestone sulphation

When OpenArt was used to visualise black crusts on limestone, the results were inconsistent like in Figure 2-4. Occasionally, the model generated accurate visualisations, suggesting potential utility, but these successes were rare.

Many images bore little resemblance to the expected effects of black crust formation, often appearing as graphic barcodes or unrelated artifacts. Additionally, as is shown in Figure6,7, OpenArt often failed to capture the entire object described, producing only partial views.

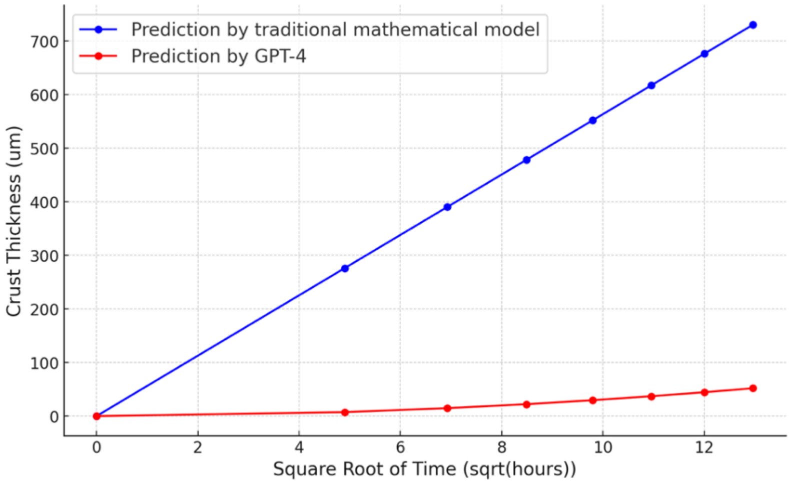

Figure 1. Comparison between mathematical predictive model and GPT-4 prediction of crust thickness over square root of time

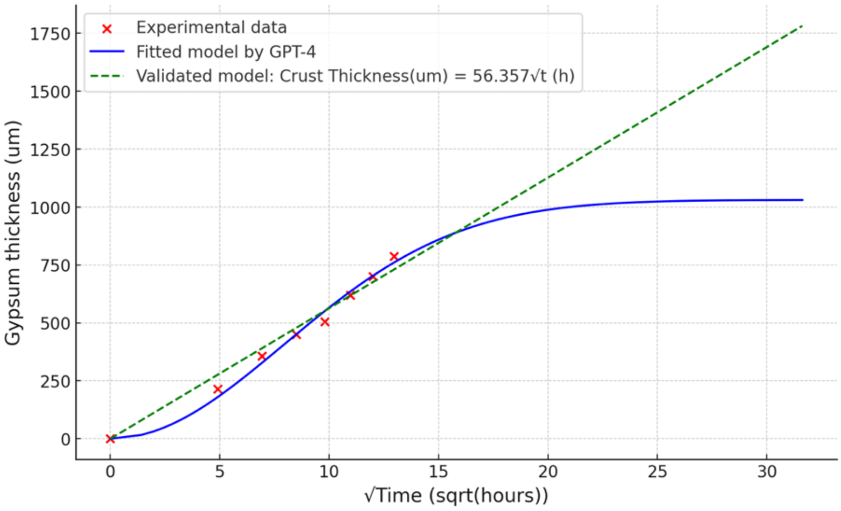

Figure 2. Comparison between mathematical predictive model and LLMs prediction of crust thickness over square root of time in 1000 hours

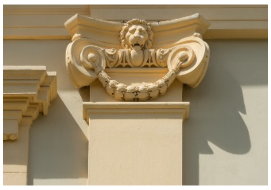

Figure 1 Original column image

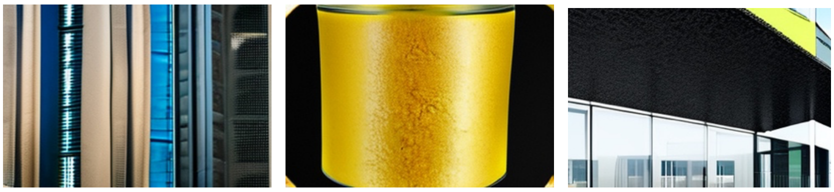

Figure 2, 3,4 Unrelated Results generated by OpenArt

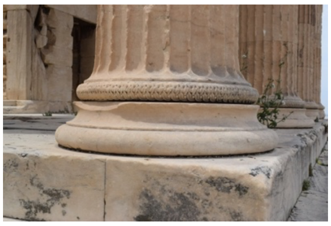

Figure 5 Original column image

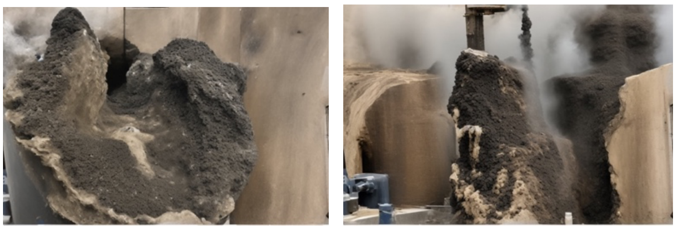

Figure 6, 7 Black crust image generated by OpenArt

4. Discussion

4.1 Unreliable and Inaccurate Output from LLMs

ChatGPT-4 can indicate general trends like increasing gypsum formation but often lacks the precision needed for scientific accuracy. GPT-4's understanding of the sulphation process is limited to simple chemical reactions, ignoring physical and environmental factors, rendering its outputs impractical. The lack of reproducibility in ChatGPT's outputs further complicates its application in heritage preservation, where consistent and accurate predictions are crucial. Claude, in contrast, provides textual responses that avoid exact figures unless strongly supported by data, reflecting a cautious approach but limiting practical applicability. OpenArt struggles with rendering complex processes like black crust formation, often producing abstract images that lack scientific detail, underscoring the challenge of applying visual AI to specialized fields like architectural preservation.

4.2 Reasons for the Unreliable and Inaccurate Output

4.2.1 Complexity of Limestone Sulphation

Limestone sulphation is a complex chemical process influenced by numerous environmental factors such as temperature, humidity, and pollution levels. The formation of black crusts involves intricate chemical reactions and physical changes that are difficult to model without a deep understanding of the underlying principles.

4.2.2 Lack of data

Effective modelling requires high-quality datasets to capture the variety of environmental conditions influencing the sulphation process. However, these datasets are limited due to the variability and complexity of microclimatic changes affecting the sulphation reaction. Similar challenges exist in using LLMs to simulate these processes, as they rely on datasets not specifically designed for heritage preservation.

4.2.3 Limitations of LLM

When generating text, LLM models often include some randomness. This is because the model usually picks randomly from a few of the most likely options for the next word or number, instead of always picking the very top option. (Spreitzer et al., 2024). And the LLM learns information during training and doesn't pull from a specific dataset. When asked the same question, it might use different patterns it has learned, picking the next word based on probability. Because of this, it can give different answers each time.(Samaan et al., 2024).

4.3 Possibilities for the Use of LLMs in Future Heritage Preservation

Despite their limitations, LLMs exhibit potential in analysing general trends in limestone sulphation. OpenArt have demonstrated the ability to generate simplistic images of black crusts, suggesting that, with further development, these models could offer valuable insights into deterioration mechanisms. This indicates that LLMs still hold potential for heritage preservation applications. Future research could include developing specialized LLM platforms tailored to chemical degradation in historical conservation. Collaborative efforts with shared data resources could improve model training and performance, leveraging expertise to tackle limestone sulphation's challenges. Expanding research could enable LLMs to become robust tools contributing effectively to cultural heritage preservation, offering accurate predictions and detailed analyses for conservation strategies.

5. Conclusion

This research assessed the potential of LLMs to predict the formation of 'black crusts' on limestone, which threatens historical structures. Comparing LLM-generated predictions with empirical data on gypsum thickness and reaction time, the study found that LLMs struggle with complex environmental conditions and subtle chemical interactions critical to limestone sulphation. Their predictions were inconsistent with traditional scientific methods, highlighting their limitations. The instability of LLMs suggests they are not yet suitable for critical decision-making in heritage preservation without support from traditional methods. An integrated approach, combining LLMs with chemical analyses, physical monitoring, and proven predictive modelling, is recommended. Future efforts should enhance LLM training datasets with detailed data from environmental monitoring, improving accuracy.

6. Reference

Alì, G., Furuholt, V., Natalini, R., Torcicollo, I., 2007a. A mathematical model of sulphite chemical aggression of limestones with high permeability. Part I. Modeling and qualitative analysis. Transp Porous Med 69, 109–122. https://doi.org/10.1007/s11242-006-9067-2

Alì, G., Furuholt, V., Natalini, R., Torcicollo, I., 2007b. A mathematical model of sulphite chemical aggression of limestones with high permeability. Part II: Numerical approximation. Transp Porous Med 69, 175–188. https://doi.org/10.1007/s11242-006-9068-1

Giavarini, C., Santarelli, M.L., Natalini, R., Freddi, F., 2008. A non-linear model of sulphation of porous stones: Numerical simulations and preliminary laboratory assessments. Journal of Cultural Heritage 9, 14–22. https://doi.org/10.1016/j.culher.2007.12.001

Giray, L., 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Ann Biomed Eng 51, 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Sabbioni, C., Zappia, G., Ghedini, N., Gobbi, G., Favoni, O., 1998. Black crusts on ancient mortars. Atmospheric Environment 32, 215–223. https://doi.org/10.1016/s1352-2310(97)00259-8

Samaan, J.S., Yeo, Y.H., Ayoub, W.S., 2024. Response to “ChatGPT’s ability to comprehend and answer cirrhosis related questions: Comment.” Arab Journal of Gastroenterology 25, 237–238. https://doi.org/10.1016/j.ajg.2024.03.002

Spreitzer, C., Straser, O., Zehetmeier, S., Maaß, K., 2024. Mathematical Modelling Abilities of Artificial Intelligence Tools: The Case of ChatGPT. Education Sciences 14, 698. https://doi.org/10.3390/educsci14070698

Tran Thi Ngoc Lan, Thi Phuong Thoa, N., Nishimura, R., Tsujino, Y., Yokoi, M., Maeda, Y., 2005. New model for the sulfation of marble by dry deposition Sheltered marble—the indicator of air pollution by sulfur dioxide. Atmospheric Environment 39, 913–920. https://doi.org/10.1016/j.atmosenv.2004.09.074